Hierarchical Task Network Planning with Common-Sense Reasoning for Multiple-People Behaviour Analysis

Hierarchical Task Network Planning with Common-Sense Reasoning for Multiple-People Behaviour Analysis

Maria J. Santofimia[a,], Jesus Martinez-del-Rincon[b], Xin Hong[b], Huiyu Zhou[b], Paul Miller[b], David Villa[a], Juan C. Lopez[a]

[a] *Computer Architecture and Networks Group, University of Castilla-La Mancha, Spain*
[b] *Centre for Secure Information Technologies, School of EEECS, Queens University Belfast, BT3 9DT, UK*

Abstract

Safety on public transport is a major concern for the relevant authorities. We address this issue by proposing an automated surveillance platform which combines data from video, infrared and pressure sensors. Data homogenisation and integration is achieved by a distributed architecture based on communication middleware that resolves interconnection issues, thereby enabling data modelling. A common-sense knowledge base models and encodes knowledge about public-transport platforms and the actions and activities of passengers. Trajectory data from passengers is modelled as a time-series of human activities. Common-sense knowledge and rules are then applied to detect inconsistencies or errors in the data interpretation. Lastly, the rationality that characterises human behaviour is also captured here through a bottom-up Hierarchical Task Network planner that, along with common-sense, corrects misinterpretations to explain passenger behaviour. The system is validated using a simulated bus saloon scenario as a case-study. Eighteen video sequences were recorded with up to six passengers. Four metrics were used to evaluate performance. The system, with an accuracy greater than 90% for each of the four metrics, was found to

Email addresses: mariajose.santofimia@uclm.es (Maria J. Santofimia), j.martinez-del-rincon@qub.ac.uk (Jesus Martinez-del-Rincon), x.hong@qub.ac.uk (Xin Hong), h.zhou@ecit.qub.ac.uk (Huiyu Zhou), p.miller@qub.ac.uk (Paul Miller), david.villa@uclm.es (David Villa), juancarlos.lopez@uclm.es (Juan C. Lopez)

outperform a rule-base system and a system containing planning alone.

Keywords: Activity recognition, common-sense, video analysis, surveillance

1. Introduction

Safety on public transportation networks is a major concern for the general public and transport authorities, specially for users and passengers as vandalism, harassment or terrorist attacks have a great impact on current society. During the last decade, there has been significant investment in the deployment of CCTV systems onboard public-transport platforms such as busses and trains (surface and underground). These systems produce enormous amounts of data that needs to be analysed in order to provide situation awareness to security analysts. However, manual analysis is not a cost-effective option. Therefore, it would be desirable if automated approaches could be developed through expert and intelligent systems.

Automating surveillance on public-transport platforms consists in recognising human activities from sensor value interpretations and video analysis. Different stages are involved in an intelligent surveillance system: detection, tracking, and behaviour analysis Gomez et al. (2015). Detection and tracking are accomplished through the use of video analytics and sensoring devices. This is, however, a challenging task because of the heterogeneity, uncertainty, and imprecision suffered by the data used for these interpretations. The fact that data has been gathered from different sources, therefore involving different devices, technologies, protocols, etc., turns data integration into yet another major challenge. Data homogenisation is therefore considered here as a previous requirement for surveillance automation. Furthermore, the scalability of the proposed solution should be assured.

On the other hand, the behaviour analysis stage faces the challenge of having to deal with erroneous, uncertain, or ambiguous data. This essential component cannot be supported on the sole analysis of video (Nebel et al., 2011), due to the complexity of the task and the fragility of current video analysis techniques.

On the contrary, artificial intelligence and statistical processing techniques are being explored as complementary sources of information that could enhance the recognition process (Edwards, 2014).

1.1. Proposed solution and contributions

Taking into account real environments where different devices, technologies, and protocols might coexist, this work starts by resolving the data integration and homogenisation problem. Moreover, if the solution proposed here is to be realistic, scalability should be seriously considered since, for an average size provincial city, a fleet can consist of several hundred busses. Our system is supported on a distributed sensor architecture, responsible for abstracting the communication issues amongst different sensor technologies and protocols. In addition, the same architecture will support the construction of advanced services for processing and fusing the information coming from sensors and the video analytics. Finally, that information will be modelled and asserted to a knowledge base where, when combined with previous and *a priori* knowledge, will derive corrections and interpolations to uncertain sensor measures.

Homogenised data enables the automation of the surveillance process. In this sense, the work in Chaaraoui et al. (2012) identifies different levels of granularity in the process of automating the task of human behaviour understanding: motion, action, activity, and behaviour. At the motion level, this work proposes the combination of different sources of information to determine the passenger motion. The tracking algorithm provides the current location of the different passengers of the scene. At the action level, video and sensor measurements have to be interpreted as actions. For example, the tracking algorithm helps on determining when the passenger is walking or the seat sensor determines when the passenger sits down or stands up. At the activity level, actions are considered as part of a greater entity, like female boarding the bus and transiting to a seat is interpreted as the activity of taking seat in a bus. Finally, at the behaviour level, actions and activities are jointly considered to explain, for example, that after boarding the bus and taking a seat the passenger has

existed the bus.

Different challenges have to be faced at each of these levels. At the motion level, sensor malfunctions or video quality or occlusions might lead to imprecise or ambiguous data. The association between this data and human motion recognition is not straight forward and further analysis is required, specially when the number of passenger increases. This paper presents a knowledge-base system to support the association process. For instance, the tracklet association algorithm relies on knowledge such as how fast a person can move inside a bus. Under ambiguous circumstances, the association process is supported on such knowledge to discard certain possible associations for being too distant one from the other, for example. However, sometimes ambiguity cannot be completely resolved using this type of knowledge. In these situations, erroneous recognition at the motion level makes the action level to be misled. To face that risk our system implements the theory of *possible worlds*. When a sensor event is suggesting that a person has sat down but the boarding sensor has not detected any passenger, it is either that the seat sensor or the boarding sensor is failing. If no further information is available both situations seem plausible so what our system does is to fork the interpretation until further information is available or it cannot be delayed any longer. At the activity level, spatio-temporal considerations have to be considered along with actions. We propose the use of common-sense knowledge to model actions, space, and time. Such knowledge combined with an appropriate reasoning mechanism will avoid situations in which, for example, a passenger has stood up from a seat before having sat down previously. Finally, at the behaviour level, activities are considered in an ordered manner. A rational approach is inspiring this work and a planning algorithm is proposed to reason at this level. In this sense, it cannot be obviated that passengers use public transport to move from one geographical point to another. In this process, they have to spend some time inside the platform transport so, whenever possible they will try to do it in the most comfortable way. Such rationality is the heuristic guiding the planning strategy proposed here to support behaviour understanding.

To summarise, this work describes a comprehensive solution to automatic surveillance in public-transport platforms, based on the analysis of multiplepeople behaviour. Several challenges have been faced, such as the heterogeneity of data and sources of information, the malfunctioning, uncertainty, and ambiguity associated to data collected from real environments, or the exponential complexity of action recognition as more than one passenger is considered in the scene. To address these challenges, several contributions are presented in this paper:

1. Development of a distributed architecture for interconnection support, data gathering and modelling.

2. Development of a comprehensive and fully automated approach for high-level semantic passenger behaviour understanding that integrates video-analytics and other sensors with a knowledge based reasoning system that deals with uncertainty.

3. Our two-stage process for multiple-people behaviour recognition and analysis supports the different levels of recognition: motion, action, activity and behaviour thanks to two novel components:

    a) A commonsensical approach for event association based on the possible worlds theory to deal with the uncertainty, vagueness, and incorrectness of context information.

    b) A Hierarchical Task Network (HTN) planning strategy to recognise courses of passenger actions, based on the aforementioned event association process.

This paper is organised as follows. First, previous work in the field of public-transport surveillance and human-action recognition are reviewed in Section 2. Section 3 describes and formalises the proposed approach for multiple-people behaviour analysis in public-transport contexts. Section 4 provides implementation details for the prototype built to experimentally validate the system.

Section 5 validates the proposed architecture. To this end, an experiment has been designed to test the proposed solution. Finally, Section 6 summarises the conclusions drawn from this work.

2. Previous work

Human behaviour recognition is a multidisciplinary field that comprises different techniques and disciplines, including machine vision, artificial intelligence and multi sensor fusion. Consequently, this sections reviews how our proposal advances current research results by analyzing the contributions made to areas such as computer vision, knowledge modeling and reasoning. This work is at the interface of computer vision and artificial intelligence where very little work have been done previously in integrating video analytics with a knowledge- base reasoning system that deals with uncertainty.

The combination of computer vision techniques with advanced mechanisms for knowledge modeling and reasoning has demonstrated a great potential for human behaviour understanding regarding approaches based solely on image and video analysis, as reported in (del Rincon et al., 2013)(Santofimia et al., 2014). However, this works consider a simple scenario in which only single-person sequences are considered. This single-person assumption is not valid for transport platforms, which are intrinsically multiple-people scenarios. These scenarios pose the problem of having to deal with associating events to *agents* since more than one passenger might have caused it. This is not a trivial task because of the precision of sensors, specially the vision sensors whose information is inherently ambiguos, and it grows in complexity as more passengers are coexisting in the scene. Our present work aims to solve this limitation by, firstly, considering and homogenizing other sources of information relevant to a transport platform (seat sensors, tracking, boarding sensor, etc.) into a distributed architecture, and by, secondly, addressing the event association problems from a novel approach that consists in seeking for a causal explanation to the sensed events.

As an intelligent system seeking a casual explanation, our work should also be compared with systems that have traditionally been intended to recognize human behaviour (Sebbak et al., 2013). In this sense, the combination of knowledge and context information has been widely studied in fields such as context-aware systems, Ambient Intelligence, Ubiquitous and Pervasive systems, etc.

Case-based reasoning approaches learn through previously acquired specific knowledge (Cocea & Magoulas, 2012)(Han et al., 2005). Consequently, they are only able to deal with situations previously presented to the system. Whilst these are the most commonly occurring, rare situations, which tend to be the most challenging and significant in our target scenario, will fail to be recognised. In contrast, the approach implemented by our work consists of general knowledge about *how the world works*, known as common sense, in addition to specific domain knowledge. As we will demonstrate, this enables the proposed system to infer and derive additional information.

As an alternative, logic-based approaches have been succesfully applied to activity recognition (Artikis et al., 2010)(Do et al., 2013). However, these approaches are normally constrained to first-order logic, which is insuficent to explain and model the human action rationality, particularly in real and multi-agent scenarios such as transport platforms. An example will be for example, a person moving from one seat to another onboard a bus due to another passenger moving closeby. In this work, we place special emphasis on modelling mental states using higher-order logic (Chen & Fahlman, 2008).

Similarly, ontological approaches, (Rodriguez et al., 2014)(Bae, 2014)(Gomez-Romero et al., 2011), are also constrained by the limitations of languages such as OWL (Bechhofer et al., 2004) or RDF (Candan et al., 2001). Neither of these allow the storing of *a-priori* inconsistent information, which is expected when multiple ambiguos sensors are employed simultaneously. Furthermore, their reasoning process is limited to consistency checking, such that no new information can be derived from existing knowledge. The solution proposed in this paper can handle *a priori* inconsistent information by postponing decision making until enough information is available, or when it cannot be delayed any

longer.

A relevant approach was presented in (Hong et al., 2016) applying probablistic reasoning, based on Dempster-Shafer (DS) theory, to address activity recognition in public transport. While probabilistic approaches can overcome the limitations of ontological languages (SanMiguel & Martinez, 2012)(Cilla et al., 2012), they require precise quantification of the uncertainty associate to every sensor and situation. In real world scenarios, such as public-transport platforms and other highly dynamic environments, deriving accurate probabilities for human activities is fraught with difficulties, since this values tend to vary over time and with the observed conditions and number of subjects (Kuipers, 1994).

To overcome the limitations of the aforementioned approaches, a commonsense reasoning strategy is proposed (McCarthy, 1968)(Minsky, 1999). Our solution has been designed to address the specific challenges of transport platforms such as the multi-agent scenario, manage inconsistent information, and exploit rich but ambiguos sensors such as cameras, which are disregarded by most previously analysed intelligent systems.

3. Methodology

This work proposes a two stage process, Figure 1, in which, based on sensor inputs, an action is first hypothesised and afterwards is reified as a concrete passenger action, performed at a specific time instant, which is part of a more general situation or an ongoing activity. The output is a set of casual explanations for each passenger's behaviour. We refer to the first stage as atomic action recognition and the second stage as situation identification.

The overall process has been conceived as a process to seek for a causal explanation to the sensed events. According to Woodaward (Woodward, 2003) a causal explanation is "*any explanation that proceeds by showing how an outcome depends (where the dependence in question is not logical or conceptual) on other variables or factors counts as causal*". Based on the supposed rationality with

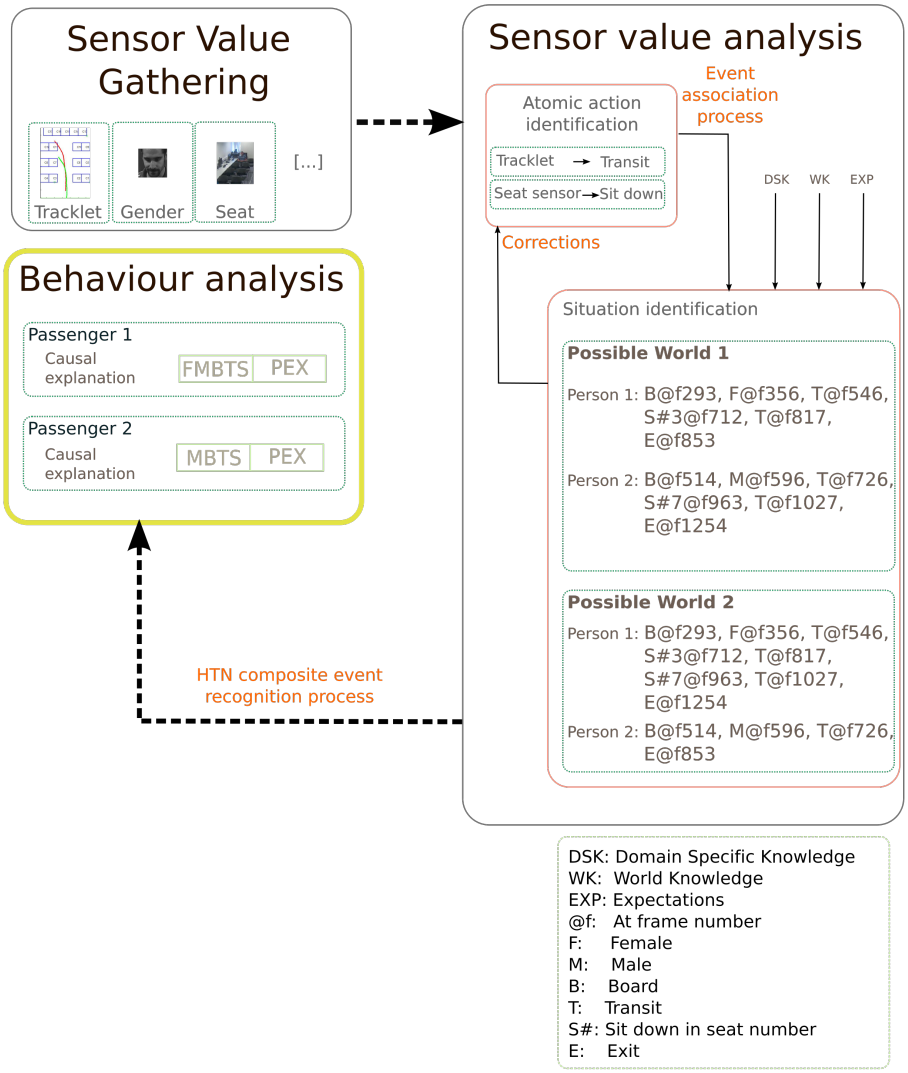

Figure 1: Overall view of the process of sensed information understanding.

which passengers behave in the public transport context, causal explanations of sensed events should match ongoing situations such as: passenger changing seat, passenger boarding bus and transiting to seat, or passenger exiting the transport platform. The first situation (passenger changing seat), for example, provides a causal explanation for events such as a seat sensor being deactivated, a tracked person tracing a route from one seat to another, and a different seat sensor being activated.

Events are not always that clear and precise due to a sensor malfunctioning or having low precision. Moreover, scenarios involving more than one person introduces an added complexity as events cannot always be unequivocally associated to a single passenger. For that reason, the first stage, Atomic Action Recognition, is simply concerned with recognising the atomic action associated to the sensed event. Then, the second stage, Situation Identification, will combine that information with Domain Specific Knowledge, World Knowledge, and Expectations. This combination will bring into light inconsistent interpretations or, alternatively, confirm the action initially hypothesised action.

More specifically, Scone[1] is used to implement the necessary mechanisms required for automating the reasoning process (Mueller, 2006) of the second stage. While other common-sense knowledge-base systems are available, such as OpenMind[2], and Cyc or OpenCyc[3], Scone was chosen for the following reasons. Firstly, OpenMind is only a database technology, lacking an inference and reasoning engine. Secondly, whilst Cyc may be more powerful, with respect to collected knowledge, it is only commercially available. OpenCyc, its open source version, is quite restricted. Finally, Scone is an open source system that provides efficient mechanisms for common-sense reasoning and knowledge modelling (Fahlman, 2006)(Fahlman, 2011). It also provides an efficient mechanism, using an abstraction called *context*, for managing *a priori* inconsistent

[1] `http://www.cs.cmu.edu/~sef/scone/`

[2] `http://openmind.hri-us.com/login.jsp`

[3] `http://www.opencyc.org/`

knowledge. The lightweight multiple-context mechanism does not overload the system even as contexts are created in the knowledge base. Moreover, the fact that only one context is active at a time means inconsistent information can be kept in the same knowledge base without causing data inconsistencies.

3.1. Multiple worlds

While the previous procedure is common in most rule-based reasoning system, our system introduces an important contribution. In a context where sensed events may have a certain degree of uncertainty, inconsistent interpretation may not be distinguishable from a coherent interpretation of inconsistent incorrectly sensed events at a given instant in time. Therefore, the recognition of inconsistent interpretations does not imply its automatic rejection. It might be that other previous interpretations, initially considered incorrect or less coherent, may be considered plausible in the light of new evidence. Since total certainty about both previous and current interpretations is not possible, inconsistent knowledge must be kept, as a backup, in case interpretations have to be re-evaluated given the appearance of new and more deterministic information. Addressing this issue however requires a knowledge-base system capable of simultaneously holding inconsistent knowledge while avoiding consistency issues. The majority of state-of-the-art knowledge-base systems (rule or ontology-based systems, just to name a few) do not provide this feature, however, Scone(Fahlman, 2010) does.

Parallel, and therefore inconsistent, causal explanations are preserved by means of an abstraction known as *possible worlds* (Divers, 2002). For example, if a seat sensor suggests that a person has sat down although no-one had previously boarded the bus, only two *worlds* can be considered possible here: one in which a person boarded the bus and the boarding sensor failed; and the other in which there is nobody in the bus and the seat sensor has malfunctioned. Each world is plausible within itself although incongruous among the others. By using this isolated world representation, inconsistent information can be represented in a logically consistent manner until future information or sensor values

will help to reduce the number of possibilities. For example, in the previous example, subsequent sensor values suggest that a passenger has stood up from the occupied chair, walks towards the exit door, and exits the bus. It can then be concluded that the only possible world is that in which the boarding sensor failed. Whereas the rest of the sensors have worked.

3.2. Atomic action recognition: Multi-target reasoning

Each passenger detected in the scene, represented as $P_i \ \forall i \in P = P_1, ..., P_n$ in Figure 2, has been associated to a set of contexts or *possible worlds*. Each context, represented by $C_j \ \forall j \in C = C_1, ..., C_s$, encompasses a set of actions $A_i \ \forall i \in A = A_1, ..., A_n$ that causally explain the detected events or so called observations. The abstract concept referred to as belief (del Rincon et al., 2013) is employed to implement each of the *contexts* or *worlds* considered plausible. A new belief will be created every time a new sensor measurement cannot be coherently, and with certainty, fitted into an existing course of action being propositionally stated in a context. Eventually, only one of these beliefs, E, the main belief, will be considered as the real estimation. Secondary beliefs hold the information considered less plausible. However, these are not discarded, being considered at first less plausible, in case further evidence reveals past choices were incorrect.

Figure 2 shows, in detail, the steps involved in the proposed process. During the association stage, sensor values have to be translated into atomic actions performed by specific passengers. Then, during the composite event recognition stage, those associations have to be processed, corrected, and interpreted such that a causal explanation can be offered that is consistent with the proposed event association.

3.2.1. Association

Association aims to establish correspondences between actions, sensor events, and actors by using spatial, temporal, and logical information. This is not trivial when more than one passenger is in the scene. In fact, the complexity exponen-

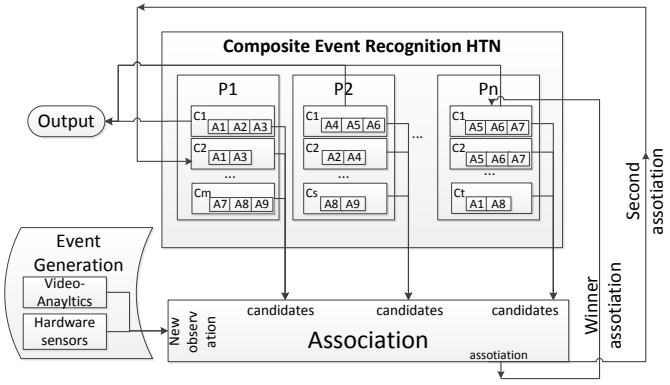

Figure 2: Event association process.

tially grows as more passengers take part in the scene, particularly when they mingle, causing events to overlap in time and space.

During the association process, the use of the aforementioned Scone mechanism for multiple-contexts (Fahlman, 2011) is essential for managing the different association possibilities. At an early stage, association decisions need to be taken in the presence of inconsistent and uncertain information, including that produced by the video analytic modules. Making hard decisions under these circumstances could lead to a misinterpretation of the scene, making it difficult to rectify later. On the contrary, managing different possibilities in different contexts allows the system to delay the association decision until new information is available, or it cannot be delayed any longer.

In conclusion, the common-sense reasoning engine, constructed using Scone, handles the association problem under temporal, spatial, and logical constraints. This engine is therefore intended to assert hypothetical atomic actions into

the most appropriate belief. Every passenger detected in the scene will have associated a set of beliefs, under which hypothetical atomic actions are being considered. Each atomic action A_i is independently processed at occurrence time t and associated to the most coherent belief b_i of the most suitable actor according to the aforementioned constraints. At time $t + \Delta$, when new atomic events are sensed, the current belief can be kept if consistent with the new information, or a new belief b_j can be created.

3.2.2. Discrete and continuous events consideration

In the public-transport scenario of interest, as in many others, some atomic events such as sitting, standing, boarding, and exiting are discrete, and occur at a specific time, while others, such as tracking events or transitions are considered continuous since they occur over a period of time. Tracking events are essential due to the rich information they potentially entail. However, among the different sensors observations considered in the bus problem, the trajectory ones are, by far, the most challenging ones.

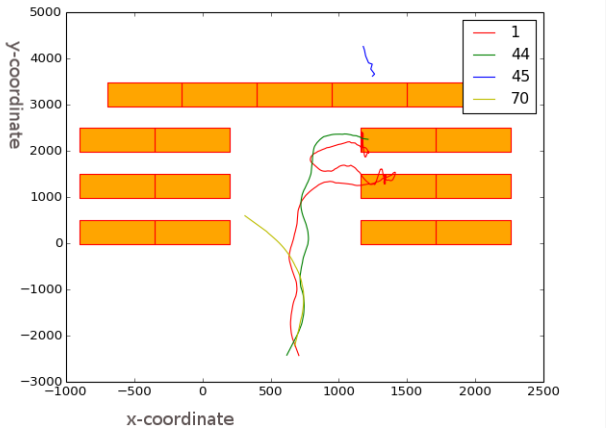

Figure 3: Trajectory representation. Orange squares represent seats and coloured lines represent trajectory fragments

It is important for the reader to understand the tracking algorithm in order to understand how the composite event recognition works. The tracker attempts

to track every subject present in the camera field of view at each time instant, regardless of their state standing/sitting down. However, it usually loses track when the actor is occluded, but it usually reacquires the actor when he/she re-appears, producing fragmented trajectories. This means that initial and end positions and times of the tracking events are not reliable indicators of atomic events. Instead, the trajectory location of actors at time instants when other discrete events happen are used during belief construction.

Figure 3 depicts the output of the video tracker for two passengers. As is clear from the figure, the task of identifying the trajectory followed by each passenger is challenging, even for humans.

Figure 4 outlines the track association process that needs to be carried out whenever a new subject detection is provided by the computer vision system. The noisy trajectory observations force the system to analyse all possible associations. Standing/sitting down detection requires special reasoning in order to corroborate the evidence from seat pressure sensors.

If a person in movement action has not been asserted into a belief in which a sitting down or exiting action exists, then it is reasonable to assume that a tracking error has occurred, therefore the reasoning engine will search for a new tracking event in the hope of linking fragmented trajectory events. This reasoning mechanism removes tracking errors produced by the video analytics, by taking advantage of the fact that more information is available from other sensors as well as the DSK/WK. For example, if two trajectory events occur that are far apart, and are separated by a large time period, based on the knowledge it has that there is only one passenger, or that all others are seated, the reasoning engine will link the events despite their temporal and spatial dissimilarity.

3.3. Situation identification: Composite event recognition

After the association stage, a set of composite events should explain the passenger behaviours, as well as determining their number. It should be expected that these composite events and resulting stories are consistent with achieving a goal that is rationally motivated (Mueller, 2007)(Wilensky, 1983)(Davidson,

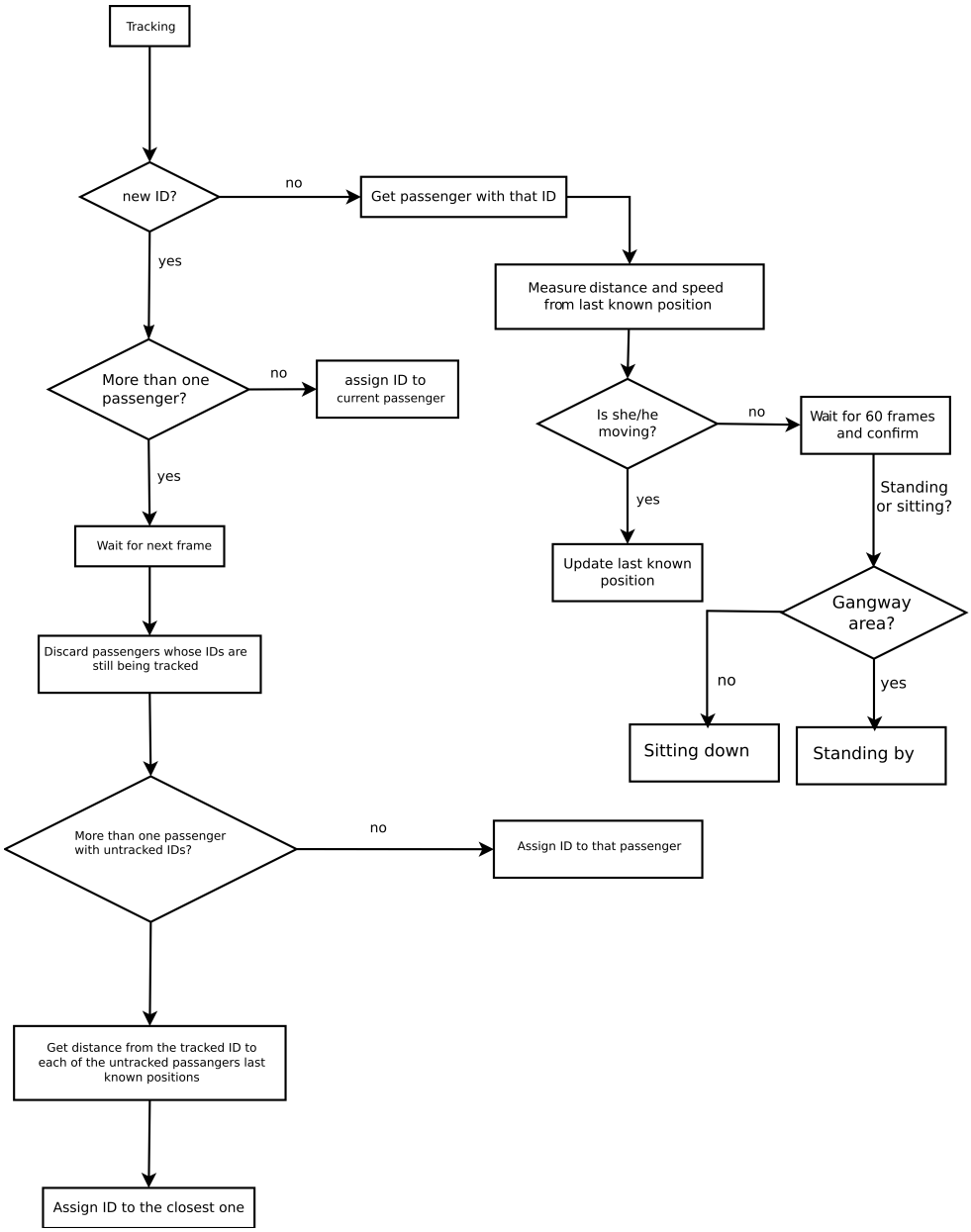

Figure 4: Tracklet association algorithm

1963). For that reason, our approach to composite event recognition will be to solve a planning problem, in which actions associated to detected events will be

consistent with a plan to achieve the ultimate goal of transporting the passenger from one point to another.

Under this approach, the task of estimating rational composite events can be automatically accomplished by adapting a bottom-up Hierarchical Task Networks (HTN) planner (Hogg et al., 2009), in which atomic actions are considered tasks.

The classic HTN algorithm is intended to determine the sequence of atomic actions that, when properly articulated, are capable of providing the functionality of composite actions. This work implements a bottom-up planner in which the knowledge retrieved from the common-sense knowledge base provides the heuristics that guide the planning algorithm.

The actions that can be performed by a passenger, at a specific location and time, are determined by his/her previous state, and the atomic actions that he/she can undertake at that location and time (pre and post-conditions in the common-sense knowledge base). For instance, if passenger is sat in seat 1, performing the action of sitting in seat 14 one second later is not feasible. Common sense tells us that to sit down in a particular seat, the person had to previously approach it, and in order to do that, the person has to be standing up and capable of transiting along the gangway.

3.3.1. Description of the planning problem

Every planning strategy has a set of common elements that define the characteristics of the problem. The first of these elements is the *state space* $\mathcal{S}$. This element describes all the states that the planner can be in. For the bus problem, the state space is determined by the description of all the possible combinations of passengers and their states (sat down, transiting, etc.). Nevertheless, due to the possibility of having infinite or very large state spaces, the proposed solution resorts to an *information space*, overlapping the state space. The information space considers the information gathered from the bus-like area sensors as well as the actions and additional observations that can be retrieved from there. For the planning problem considered here, states are depicted as a tuple of the form

$s = (P_i, A_i)$ meaning that the current situation $s \in \mathcal{S}$ is the resulting state after a certain passenger $P_i \in \mathcal{P}$ performs the action $A_i \in \mathcal{A}$.

The second element of a planning problem consists of the set of actions or *action space* $\mathcal{A}(s)$ available at each given state $s \in \mathcal{S}$. Availability is determined by the satisfaction of the pre and post-conditions of each of the considered actions. Action unavailability means that a certain action either leads the system to a situation incompatible with the desired goal state or requires a different state of the world to take place.

An additional element is the *state transition function* f that, given the current state and action space, produces a new state for every action, out of the action space, that is currently available.

$f(s, a, \theta)$ for $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $\theta \in \Theta(s, a)$

Function f returns the actions, from the action space, that are available in the current situation s.

The Θ function therefore provides the set of situations that can be reached given the current state and the execution of any of the actions that are available at that state.

Stages, denoted by $k \in \mathcal{K}$, also need to be considered in order to conceive the execution plan as an incremental task. Moreover, stages are used by the planner to evaluate the evolution of the plan execution, identifying possible deviations from it. In this incremental context, the *goal state* is specially relevant, denoted by $\mathcal{S}_G \subset \mathcal{S}$. The goal state in our scenarios will be determined by the passenger existing the bus, due to two main reasons. First, all the passengers are supposed to exit the bus eventually. Second, since no more information is expected to be received following exit, the reasoning system must take a final decision regarding the composite events performed.

Finally, it is necessary to have a function that evaluates the *goodness* of an action in comparison with the others. Given the current and the goal situation, the cost function L weights each action according to its suitability in seeking the course of actions that minimises the cost of reaching the goal state. The history of states, actions, and available actions are respectively denoted by $\tilde{s}_K, \tilde{a}_K, \tilde{v}_K$,

so the cost of a given course of actions, given that G is the goal state, can be calculated as follows:

$$L(\tilde{s}_K, \tilde{a}_K, \tilde{v}_K) = \sum_{k=1}^{K} l(s_k, a_k, v_k) + l_G(s_G) \tag{1}$$

The cost function should consider a set of constraints l involved in the process of action association:

1. Temporal constraint T_c: Atomic events are temporal entities and beliefs are created as consecutive sequences of events (atomic actions).

2. Spatial constraints D_c: The further apart two atomic events happen, the less likely is that they are associated. The distance between the last known locations of the actors (in case of tracking events). or sensor locations according to DSK (in case of boarding, exiting or sitting down events) are computed.

3. Common-sense constraints G_c: Defined in the WK. This constraint is binary since it deals with impossible incoherence such as people cannot sit down if they are already sat down.

These constraints can be modelled as a cost function l

$$l(a_i, p_j, \theta_k) = \begin{cases} D_c(p_j, a_i) & if \quad G_c(p_j, a_i) * T_c(p_j, a_i) == true \\ \infty & if \quad G_c(p_j, a_i) * T_c(p_j, a_i) == false \end{cases} \tag{2}$$

where j is the index of the possible actors that can perform the given action a_i, and D is the Euclidean distance between the locations of the atomic events to be associated. Common-sense constraints G_c and temporal constraints T_c of passenger p_j performing action a_i are implemented as gating functions, where an impossible association is assigned an infinite cost.

3.3.2. The planning algorithm

The planning strategy considers all previous elements in order to achieve the goal state. Our proposed planning algorithm uses a Dijkstra-like algorithm

(Dijkstra, 1959) in which each stage of the execution of the plan is expected to be closer to the goal than the previous stage.

In contrast to the traditional Dijkstra algorithm, states cannot be weighted beforehand, but are evaluated at each execution stage, based on the cost function in eq. 2. Moreover, revisiting states is allowed since states are described in terms of the situation that results from a passenger performing an action.

Algorithm 1 HTN planning(s_0, s_g)

$\Pi = (\mathcal{P}, \mathcal{A})$
for every passenger p_c in $\mathcal{P}$ **do**
  $s_0 = (p_c, a_0)$ that have arisen the goal $s_g = (p_c, a_g)$
  $s_c = s_0$ and $a_c = a_0$ current values are the initial values
  while s_c is different from s_g **do**
    get all the actions a_i that are available in the current state s_c
    $f(s_c, a_c, \theta_c) = (a_0, a_1, a_2, ...)$
    $\theta = \Theta(p_c, a_i)$
    for θ_i **do**
      get its cost function $l_i = L(s_c, a_i, \theta_i)$
    end for
    select the action that minimises the cost function $L\ MIN(l_i)$
    Append p_c, a_i as π_i
    $s_c = \Theta(g_c, a_i)$
  end while
end for
Return Π

The planning algorithm starts from situation s_0 that differs from the goal state s_g that results from the passenger p_c performing the action a_i exiting the bus. The HTN planning algorithm is devised to find the sequence of actions $\mathcal{A}$, performed by every passenger comprising the actor set $\mathcal{P}$ considered in the scene.

For every passenger p_c from the list of passengers in the scene $\mathcal{P}$, while the

current situation does not match the goal situation, the algorithm will look for the action that moves the current situation closer to the goal state. Therefore, given a current situation, the list of possible actions is determined by the function $f(s_c, a_c, \theta_c)$. It is possible to rank all the possible actions, based on the spatio-temporal and common-sense constraints, using the cost function given by Equation 2. The action that minimises the cost will be selected as the action carried out by the passenger p_c leading the context to situation s_c, which is a step closer to the goal state s_g.

4. System implementation

Previous section has described the theoretical and technical details of the proposed methodology for ongoing-situation identification. Despite being the key module of a system aimed at public-transport surveillance, this module has to be supported by some others for information gathering, communication support, and knowledge modelling. This section therefore provides implementation details for these other modules involved in the proposed solution.

Figure 5 provides a system overview depicting the different stages involved in the process of situation understanding. The first step consists in gathering information from the bus-like area sensors and the video analytics. It has to be noticed that, despite the fact that the implemented prototype works upon prerecorded sequences, from the point of view of the middleware abstraction layer, there is no difference on whether sensor measures are being published by real or post-processed sensors. In this sense, measure will be published in a distributed communication channel. Then, the reasoning system subscribed to that communication channel will be notified whenever a new publications appears in the channel. Whether these publications are the result of a real sensor or a synthetic service make no difference for the reasoning system.

The method described in Section 3 will come into play for every new sensor publication, to firstly hypothesise the atomic action that will afterwards be justified or corrected based on the available knowledge.

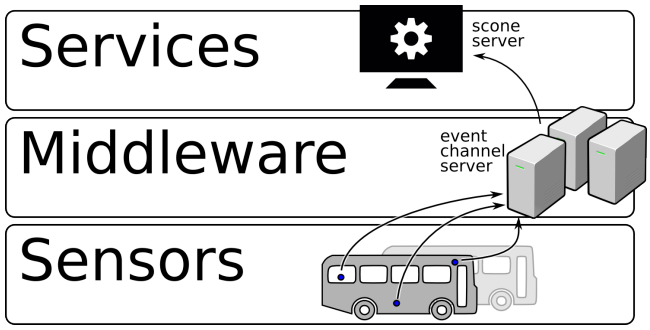

Figure 5: System overview

4.1. The distributed architecture

Sensorized contexts such as the bus one, are characterized by the presence of different hardware devices that use different protocols, operating systems, or implementation languages. Collecting information from these devices implies the implementation of a distributed heterogeneous application. In this sense, the use of a middleware technology simplifies the procedures required for achieving an effective communication among the different devices involved in the application.

The solution proposed here resorts to ZeroC ICE[4] as the commercial distributed object-oriented middleware technology. ZeroC ICE is an object-oriented and CORBA-like middleware technology that provides the means (tools, API, libraries) to easily build object-oriented client-server applications. Despite being similar in concept to CORBA, there are some additional resources that make ZeroC ICE the most appropriate technology for the solution devised here. In this sense, two of the most useful services provided by this technology, `IceGrid` and `IceStorm` play an essential role in easing the application deployment as well as in abstracting the details involved in implementing a publish/subscribe architecture.

Regarding scalability, the ZeroC ICE technology provides an implementation

[4]http://www.zeroc.com/

of the evictor pattern, as well as mechanisms to automate object persistence, that ensure the scalability of the system.

4.2. The computer vision system

Sensor data and video analytics are fed into the reasoning system in order to provide the require information to discover the underlying actions and behaviours. As mentioned before, our system combines hardware sensor inputs, when it is possible to install them within the transport platform without being intrusive or prohibitively costly, with video sensors, given their highly potentially rich information and their low intrusiveness. However, in spite of these benefits video sensors by themselves only provide raw pixel information, which is not of much use for an automatic reasoning system. In order to fully exploit video sensors, computer vision algorithms are being used to extract automatically relevant information.

The first video analytics module is a gender recognition system. A camera pointing at the entrance/exit of the transport platform provides the data to our recognition algorithm. This system is composed of 3 different component as depicted in Figure 6. Once a new image is capture, the first component detects the passenger's face by applying Viola and Jones face detector (Viola & Jones, 2014). In this detector a bank of rectangular Haar filters are used to extract contrast features from the image and then feed into a boosting cascade classifier to label the image region as a face or not. By fully scanning the image horizontally and vertically using a sliding window the presence and absence of faces as well as their location is determined.

After locating the face, its corresponding pixels are first projected onto a reduced subspace derived using a principal component analysis (PCA), aiming to reduce the dimensionality by discarding irrelevant and noise information within the face image (several hundreds of pixels even in low resolution). Then, the resulting reduced feature vector is input to a support vector machine (SVM), which classifies the image as male or female by finding the separating hyperplane with the maximal margin between both populations. Both PCA and SVM

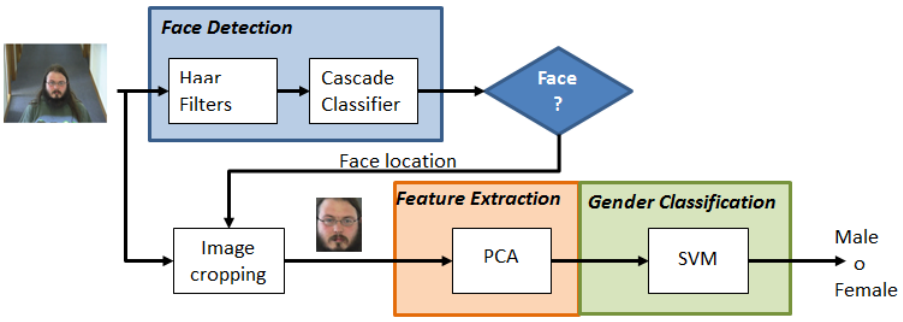

Figure 6: Tracklet association algorithm

requires a training data set of face images, which is composed of 1841 female and 1918 male face images. The resulting output of our gender recognition system is a label with the gender of the passenger as well as a confidence value or probability of the face as being either male or female. The accuracy of this video analytic module was reported to be 83% in an independent testing (Stewart et al., 2009).

The second video analytics module is a multi-target tracking system, shown in Figure 7. A second camera pointed along the bus saloon aims to capture the movement of the passengers within the saloon. The tracking-by-detection algorithm consists of four stages. Firstly, a Poselet detector (Bourdev & Malik, 2009) is applied to detect signatures of humans in the video on a frame-by-frame basis. Secondly, a calibrated process is used to project the detections from the image plane into the 3D real space and also discard those detections which are likely to be false positives, e.g. people of abnormal size or detections that are located outside the bus. This calibration process also allows us to know the 3D locations of all seats, the gangway and the entrance/exit, which can be later correlated with the passenger position for further reasoning.

Finally, human detections are linked together over time using a hierarchical dual-stage linear assignment procedure to form tracks of the passengers. In the first stage, detections are associated on a frame-to-frame basis by using

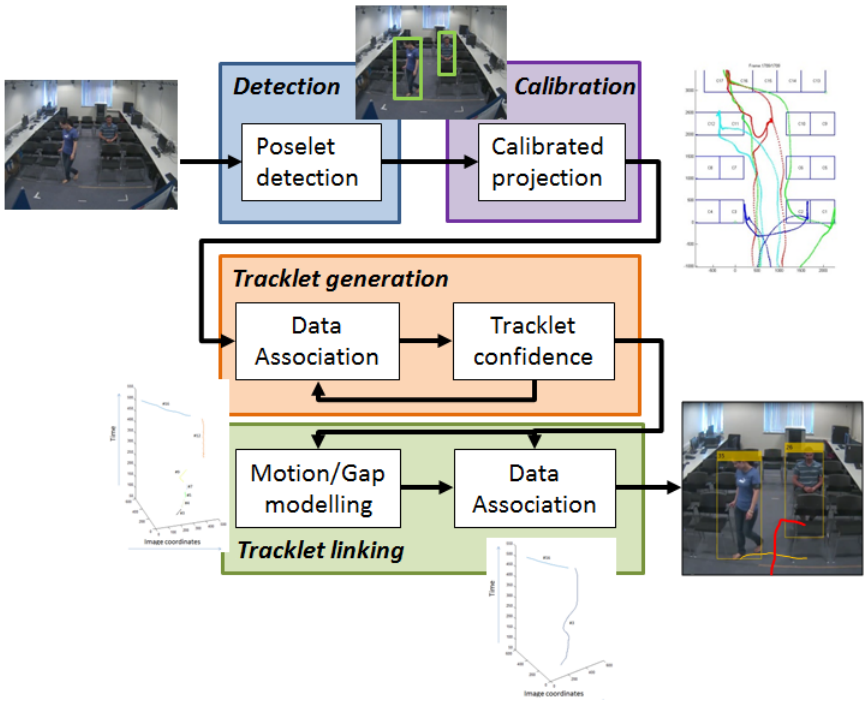

Figure 7: Tracklet association algorithm

their colour appearance, temporal locality and spatial distance. The resulting fragmented tracklets, are subsequently linked into longer tracks by a second level of linear assignment, where reasoning about the gaps and the interactions with other passengers can be modelled. A full description of this tracking algorithm is provided in (McLaughlin et al., 2014) and details about the tracklet confidence are provided in section 5.2. The final output of the module is a set of trajectories, containing the passenger 3D location and identity (sequentially allocated labels) at every temporal instant.

In addition to the video sensor and analytics, two other hardware sensor types are simulated using a VICON tracking system (Ltd, 1984); a pressure sensor at every seat, to detect when a passenger sits down/stands up, and an infrared motion detector, to detect when a passenger gets in/out, at the entrance/exit of the bus.

4.3. Sources of knowledge

To validate our working hypothesis, a common-sense system has to be built so as to model and reason about the information obtained from sensors, video analytic modules, the context and the world itself. In this sense, the automation of the reasoning task requires a language and a syntax, a knowledge base comprising the available information and rules, and a consistency checking mechanism that makes use of the available knowledge base and information provided by the sensors to infer new coherent information. Our current common-sense framework has been implemented using Scone (Fahlman, 2006) due to its suitability for modelling actions and human behaviour. By using Scone, it is possible to encode, using a LISP-like syntax, formal definitions describing the World knowledge (WK) and Domain specific knowledge (DSK), as well as the expected set of behaviors, here referred as expectations (EXP).

These three sources of knowledge are described below:

1. World knowledge, WK, comprises all relevant common-sense knowledge that describes "*how the world works*". This information is independent of the application domain or any particular scenario. It only considers general knowledge rather than specific or expert knowledge. As an example, we provide below the description of the action of 'boarding a bus'.

Listing 1: Boarding action

```
1
2 (new-action-type {boarding}
3                  :agent-type {passenger}
4                  :object-type {movable object})
5
6 (new-action-type {boarding bus}
7                  :agent-type {passenger}
8                  :object-type {bus})
9
10 (new-is-a {boarding bus} {boarding})
11
12 (new-context {boarding bus BC} {general})
13 (new-is-a {boarding bus BC} {before context})
```

```
(x-is-the-y-of-z {boarding bus BC} {before context} {boarding bus})

(new-context {boarding bus AC} {general})
(new-is-a {boarding bus AC} {after context})
(x-is-the-y-of-z {boarding bus AC} {after context} {boarding bus})

(in-context {boarding bus BC})
(new-statement {passenger} {approaches} {bus gate})
(new-not-statement {passenger} {passes through} {bus gate}))
(new-statement {passenger} {stands on} {land})
(new-not-statement {passenger} {is in} {bus})

(in-context {boarding bus AC})
(new-statement {passenger} {stands on} {bus floor})
(new-statement {passenger} {passes through} {bus gate}))
(new-statement {passenger} {is in} {bus})
```

2. Domain specific knowledge, DSK, describes a given application domain in terms of the entities that are relevant for that specific context, as well as, the relationships established between them. The description of sensor placements or the seat distribution, as part of a specific bus layout, are examples of DSK. Listing 2 provides a description of the coordinates (in centimeters) with respect to the camera perspective.

Listing 2: Bus specific knowledge

```
(new-type-role {x-coord} {position} {location})
(new-type-role {y-coord} {position} {location})

(new-type {bus entrance position} {position})
(new-type {bus chair} {static object})
(new-type-role {bus chair location} {bus chair} {position} :n 8)
(new-indv {seat 1} {bus chair})
(x-is-a-y-of-z {2260} {x-coord} {seat 1})
(x-is-a-y-of-z {-20} {y-coord} {seat 1})
(x-is-a-y-of-z {1710} {x-coord} {seat 1})
(x-is-a-y-of-z {-20} {y-coord} {seat 1})
(x-is-a-y-of-z {1710} {x-coord} {seat 1})
(x-is-a-y-of-z {500} {y-coord} {seat 1})
```

```
(x-is-a-y-of-z {2260} {x-coord} {seat 1})
(x-is-a-y-of-z {500} {y-coord} {seat 1})

(new-type {passenger} {person})
(new-type {bus passenger} {passenger})
(new-type-role {bus passenger position} {bus passenger} {position})

(new-type {sensor} {thing})
(new-type {infrared barrier} {sensor})
(new-type-role {infrared barrier location} {infrared barrier} {
    location})
(new-statement {infrared barrier} {is in} {bus})
(new-statement {infrared barrier} {controls} {bus gate})
```

3. Expectations, EXP, consist of sequences of actions that are expected to occur. It encapsulates logical concepts such as causality, motivation, and rationality, which are expected in human action recognition, in particular for passengers onboard. For example, in a bus context, if a person boards the bus, that passenger is expected to walk along the aisle and sit down if seats are available (Listing 3). Expectations are part of the domain specific knowledge since the described behavioural patterns are context-specific. Different behaviour of the same passenger could be expected in a different transport platform, such as a train or airplane, where seats are pre-allocated.

Listing 3: Bus specific knowledge

```
(new-type {expectation} {thing})
(new-type-role {has expectation} {expectation} {event})

(new-indv {MBTSt} {expectation})

(x-is-the-y-of-z  {male boards} {has expectation} {MBTSt})
(x-is-the-y-of-z {male transists to} {has expectation} {MBTSt} )
(x-is-the-y-of-z {male sits} {has expectation} {MBTSt} )
(the-x-of-y-is-a-z {action agent} {male boards}  {male})
(the-x-of-y-is-a-z {action agent} {male transits to}  {male})
(the-x-of-y-is-a-z {action agent} {male sits}  {male})
```

4. Beliefs, BLF, consist of a mechanism which attempts to replicate a hu-

man's ability to recognize actions in poor quality video, or other sensor that provide ambiguous information. If there is only one active expectation, the belief will trust and follow it. However, if, due to sensor ambiguity, multiple expectations are active, the belief will select the most appropriate one to assert the action happening in the scenario. This mechanism is implemented in Scone through the multiple-context mechanism.

5. System evaluation

5.1. The dataset

In order to validate our approach, a bus saloon scenario was simulated within a laboratory. The setup was designed to resemble a bus saloon as much as possible. It includes an entrance/exit doorway, a gangway and two parallel seated areas plus a full row at the end, giving a total of 17 seats (C1-C17), Figure 8.

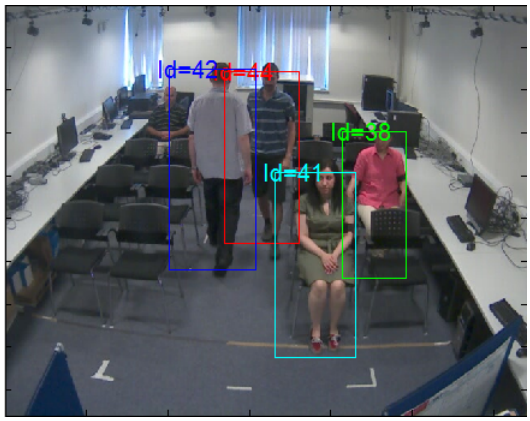

Figure 8: Recreation of bus saloon

Two cameras were mounted in the lab with a similar elevation and tilt to those onboard a real bus. The first camera is placed to capture the bus entrance and to easily facilitate face detection and gender recognition. The second camera is located to capture the bus saloon area and to facilitate tracking of passengers as they transit to and from seated areas.

The following **sensor events** obtained from the dataset:

- Entrance detector.
- Gender classification.
- Trajectory fragments, also known as tracklets.
- Pressure sensor detecting sitting down.
- Pressure sensor detecting standing up.
- Exit detector.

Each of these sensor events has a unique *sID* (the tracklets sID is initially the one given by the tracker and the other sensors have a time-stamped ID provided by the communication-channel publisher).

Recall that an **atomic event** is considered here as a short sequence of sensor events with limited purpose or intent. Each atomic event has also a unique label *aID*. In our scenario the following atomic events are considered:

- **Person boarding**: entrance detector plus gender classification. At this point a person ID *pID* is assigned to the passenger.
- **Person transiting**: one or more tacklets (or even part of a tracklet).
- **Person sitting down**: Pressure sensor detecting sitting down.
- **Person standing**: Pressure sensor being released.
- **Person Exiting**: Exit detector.

Composite events have been defined here as sequences of atomic events (or longer sequences of sensor events) with a rational purpose in the context of a public-transport scenario. Similarly, each of these composite events has a unique identity *cID*. The following composite events have been considered in our scenario:

- **MBTS/FMBTS**: Male or Female boarding the bus and transiting to a seat. Composed of: Person pID(i) boarding + person pID(i) transiting + Person pID(i) sitting down.

- **PCS**: Person changing seats. Composed of: Person pID(i) standing up + person pID(i) transiting + Person pID(i) sitting down.

- **PEX**: Person exiting the bus. Composed of: Person pID(i) standing up + person pID(i) transiting + Person pID(i) Exiting

Finally, **stories** or **situations** are considered here as sequences of composite events (or a larger sequence of sensor events) having a unique *pID* identity, i.e. the full sequence of events of a given passenger from the moment they board to the moment they exit the bus .

Six different subjects, three males (M) and three females (F) took part in the capture of this validation dataset. A total number of eighteen sequences of varying complexity were recorded. Table 1 summarises the actors, actions and behaviours occurring in each sequence, whilst Figure 9 depicts the seat distribution in the considered scenario.

Table 1: Description of the dataset sequences. @f indicates the frame number at which the composite event starts.

Scene	**Actors**	**Frames**	**Composite Actions**
DL1 ACT3 01	1 M, 1 F	1071	MBTS-C14 @f20, FBTS-C6 @f298, MEX @f665, FEX @f953
DL1 ACT3 02	1 M, 1 F	1005	MBTS-C15 @f25, FBTS-C7 @f376,MEX @f556, FEX @f896
DL1 ACT3 03	1 M, 1 F	960	MBTS-C15 @f20, FBTS-C8 @f252, MEX @f629, FEX @f819
DL2 ACT2 01	1 M, 1 F	1367	MBTS-C17 @f23, FBTS-C7 @f338,MCS-C9 @f626, MEX @f991, FEX @f1251

DL2 ACT2 02	1 M, 1 F	1314	MBTS-C10 @f28, FBTS-C7 @f432,MCS-C14 @f625, MEX @f919, FEX @f1203
DL2 ACT2 03	1 M, 1 F	1343	MBTS-C13 @f21, FBTS-C7 @f288,MCS-C6 @f624, MEX @f1071, FEX @f1226
DL2 ACT4 01	1 M, 1 F	1165	MBTS-C7 @f19, FBTS-C7 @f332, MEX @f643, FEX @f1046
DL2 ACT4 02	1 M, 1 F	933	MBTS-C10 @f27, FBTS-C10 @f354, MEX @f551, FEX @f805
DL2 ACT4 03	1 M, 1 F	835	MBTS-C7 @f22, FBTS-C7 @f216, MEX @f391, FEX @f726
DL2 ACT3 01	1 M, 1 F	1134	FBTS-C5 @f15, MBTS-C6 @f267, FCS-C3 @f492, FEX @f914, MEX @f1019
DL2 ACT3 02	1 M, 1 F	967	FBTS-C5 @f16, MBTS-C6 @f224, FCS-C3 @f380, FEX @f792, MEX @f847
DL2 ACT3 03	1 M, 1 F	913	FBTS-C2 @f17, MBTS-C1 @f310, FCS-C7 @f448, FEX @f749, MEX @f828
DL3 ACT01	1 M, 1 F	1405	FBTS-C6 @f17, MBTS-C10 @f265, FCS-C3 @f553, MCS-C7 @f762,FEX @f1075, MEX @f1294
DL3 ACT02	1 M, 1 F	1279	FBTS-C6 @f20, MBTS-C11 @f292, FCS-C11 @f618, MCS-C6 @f651, MEX @f1011, FEX @f1167
DL4 ACT02	1 M, 1 F	870	MBTS-C15 @f23, MEX @f375, FBTS-C6 @f449, FEX @f763
DL4 ACT01	2 M, 1 F	1657	M1BTS-C6 @f21, FBTS-C6 @f289, M1CS-C8 @f558, M2BTS-C6 @f818, FEX @f1067, M1EX @f1346, M2EX @f1544

DL5 ACT02	2 M, 2 F	1789	M1BTS-C17 @f18, F1BTS-C12 @f294, M2BTS-C17 @f514, M1CS-C1 @f728, F2BTS-C2 @f996, F2CS-C3 @f1168, M1EX @f1354, M2EX @f1456, F1EX @f1542, F2EX @f1712
DL5 ACT01	3 M, 3 F	1432	F1BTS-C7 @f15, M1BTS-C14 @f179, F2BTS-C5 @f308, F1EX @f494, M2BTS-C17 @f442, F3BTS-C3 @f696, M1EX @f818, M3BTS-C13 @f799, M2EX @f968, F2EX @f1029, F3EX @f1235, M3EX @f1274

The first twelve sequences aim to represent a spectrum of possibly risky behaviour patterns. The goal in these sequences is to explore the potential application of our event recognition system for future automatic risk assessment within a transport scenario. Sequences DL1 ACT3 simulate a normal bus journey (zero risk situation) where a couple of passengers undertake their trip without interacting. Sequences DL2 ACT2 simulate a low risk situation where a passenger changes seat whilst the bus is moving. This may be indicative of a passenger who may feel threatened, or one who is trying to threaten another passenger. Sequences DL2 ACT4 simulates a medium risk situation, where a passenger loiters near another who is sitting down. Sequences DL2 ACT3 simulate a high risk situation, where a male passenger sits beside a female passenger who immediately moves.

The last six sequences include increasingly complex scenarios with more passengers and greater interaction between them. Changing seats (DL3), crossing in the gangway (DL4 ATC02), and multiple interactions between multiple passengers are recurrent situations. The goal of these sequences is to evaluate the upper limit of events and actors that our system is able to recognise.

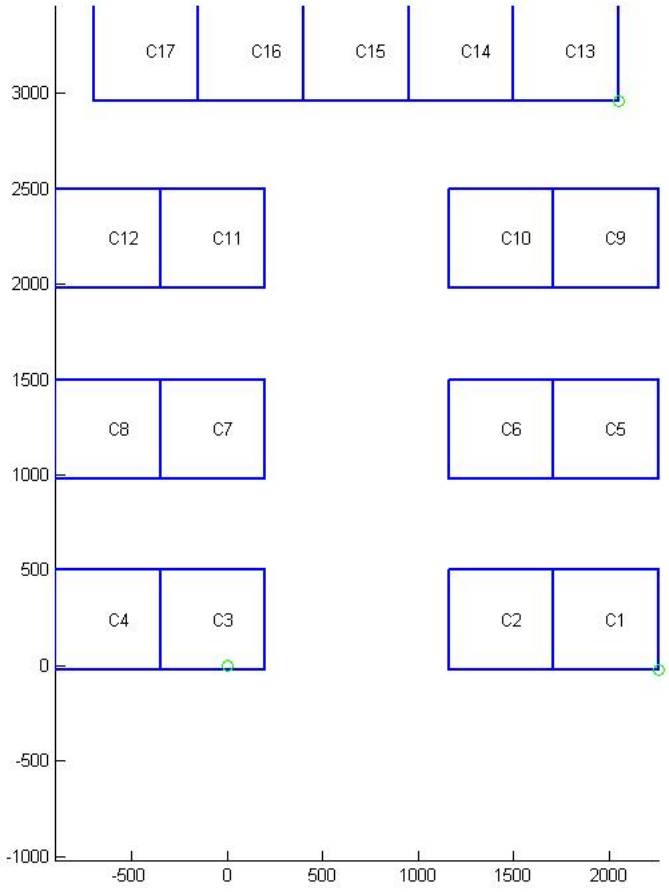

Figure 9: Bus seat distribution

5.2. Video analytics performace

In order to fully evaluate our proposed reasoning system, and since real sensors were used, it is important to evaluate the performance of the video analytics so a better understanding of the capabilities of the reasoning engine is achieved. While perfect sensors and analytics will mean that the reasoning does not need to be specially robust, imperfect sensors require better reasoning to address real world problems. Obtaining a high recognition rate of the actions and behaviours given imperfect sensors, demonstrate the potential of the reasoning system to

correct errors and work with uncertainty.

In the described dataset, a total number of 43 passengers appear in the different sequences. The accuracy *Acc*, or recognition rate, of the gender classifier used, defined as the number of boarding passengers whose gender is correctly classified divided by the total number of boarding passengers, is:

$$Acc = \frac{37}{43} = 0.86 \tag{3}$$

This number, 86%, is very similar to that reported by the authors in their paper 83% (Stewart et al., 2009). Although the performance of this classifier is reasonably high, it must be noticed that, since it is the only sensor evaluating gender, an error in the gender recognition cannot be corrected by further reasoning.

The evaluation of the tracking system is more complex since, instead of binary errors as in the gender system, multiple types of errors can occur. In order to evaluate the performance of the algorithm, a combination of Type I and Type II errors -true positive, true negative, false positive, false negative- and multi target specific metrics -identity swaps, MOTA, etc.- must be used. A detailed explanation of the metrics is displayed in Table 2, and the quantitative performance of the tracking system in Table 3. More details about how these metrics are calculated and evaluated can be found in (Bernardin & Stiefelhagen, 2008).

One can observe that the tracking, despite being state-of-the-art, is far from providing perfect results, with only 25% of trajectories tracked, a high number of missing frames (FN), fragmented tracklets, and a middling value for recall and MOTA. Given the high percentage of errors, and the tracklet fragmentation, simple association of events may not be enough to solve complex sequences. However, when the target is successfully tracked by the system, the precision is reasonably good and the sensor can be trusted, as evidenced by the high recall and MOTP.

5.3. Experimental setup

Four metrics are considered in order to evaluate the system at different levels.

Table 2: Metrics used to evaluate quantitatively a tracking system

Measure	**Better**	**Perfect**	**Description**
GT	-	-	number of ground truth trajectories
MT	higher	GT	Mostly tracked targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.
PT	-	-	Partially tracked targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 20% and at most 80% of their respective life span.
ML	lower	0	Mostly lost targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.
FP	lower	0	The total number of false positives
FN	lower	0	The total number of missed targets or false negatives
ID Sw.	lower	0	The total number of identity switches. Please note that we follow the stricter definition of identity switches as described in (Li et al., 2009)
FM	lower	0	The total number of times a trajectory is fragmented (i.e. interrupted during tracking)
Recall	higher	100%	Percentage of detected targets
Precision	higher	100%	Percentage of correctly detected targets
FAR	lower	0	The average number of false alarms per frame.
MOTA	higher	100%	Multiple Object Tracking Accuracy (Bernardin & Stiefelhagen, 2008). This measure combines three error sources: false positives, missed targets and identity switches
MOTP	higher	100%	Multiple Object Tracking Precision (Bernardin & Stiefelhagen, 2008). The misalignment between the annotated and the predicted bounding boxes.

Table 3: Quantitative performance of tracker system on our dataset

Performance							
GT	43	FP	5	Recall	56.3	MOTA	56.1
MT	11	FN	16538	Precision	100.0	MOTP	68.2
PT	28	ID Sw.	90	FAR	0.00		
ML	4	FM	89				

- **Metric 1. Sensor association accuracy**: assessing the number of *sID* correctly associated to its pID divided by the total number of *sID*.
- **Metric 2. Atomic event association accuracy**: assessing the number of *aID* correctly associated to its *pID* divided by the total number of *aID*.
- **Metric 3. Composite event association accuracy**: assessing the number of *cID* correctly associated to its pID divided by the total number of *cID*.
- **Metric 4. Story recognition accuracy**: assessing the number of stories correctly composed divided by the total number of *pID*.

Metric 1 measures if the sensor events generated by the system are correctly associated to the person that triggered them. Since only generated events are considered, sensor errors such as missed sensor events are not included in this metric. Metrics 2, 3 and 4 are compared against the manually annotated groundtruth, so missing events and sensor errors are considered and expected to be corrected by the reasoning engine. Metric 2 considers all atomic events, including those that cannot be improved with common sense reasoning, such as the gender classification. Therefore, this aspect has been obviated for metrics 3 and 4. Metric 4 is the most significant one given the fact that it is the desired outcome of the system and that a single error or atomic event mistake can invalidate the full story.

5.4. Results

Table 4 summarises the accuracy obtained by the proposed system with respect to each of the aforementioned metrics.

Regarding metric 1, it can be observed that in 14 out of the 16 sequences, involving 2 or 3 passengers, sensor events were correctly associated to each of the passengers involved. However, we can point to seat proximity as the reasons why events in the other two sequences, were incorrectly associated. For example, in sequence DL2 ACT2 02, the IDs association to passengers fails because the passengers were sat too close to each other, in seats 10 and 14, making tracking and association difficult. Another possible reason why the system might fail to associate passengers with the correct ID is when the tracking struggles to detect a particular person in the image (due to clothes, light, orientation, etc.). In cases such as the sequence DL2 ACT3 03, the tracking system output was so poor, that the reasoning system could not make any improvement at the sensor association level. Regarding scenarios where more than two or three passengers are involved, the accuracy rate drops, as can be observed for sequences DL5. This is due to the aforementioned problems experienced by the tracking system.

Regarding metric 2, the reasoning engine is able to address the mistakes at sensor level and recognise most of the atomic events, providing an accuracy greater than the one obtained for metric 1. The most common remaining errors are due to the gender recognition, which means that there is little the reasoning system can do to correct that situation.

The bigger picture is analysed by metrics 3 and 4. Regarding these two metrics, it can be concluded that most of the complex events, as well as the full stories for each of the passengers from the moment they board to the moment they leave the bus, are correctly recognised, despite multiple errors at lower levels. This is due to the fact that the reasoning engine and the different mechanisms, such as multiple context or HTN, have more information available at those levels with which to reason correctly. The performance is above 75% in all the cases, even for those sequences involving six passengers with multiple interactions. The full description of the reconstructed sequences is provided in

Table 4: Full system accuracy rates obtained for each sequence and metric

System	**Metric 1 (%)**	**Metric 2 (%)**	**Metric 3 (%)**	**Metric 4 (%)**
DL1 ACT3 01	100.0	96.77	100.0	100.0
DL1 ACT3 02	100.0	100.0	100.0	100.0
DL1 ACT3 03	100.0	100.0	100.0	100.0
DL2 ACT2 01	100.0	100.0	100.0	100.0
DL2 ACT2 02	75.0	97.22	100.0	100.0
DL2 ACT2 03	100.0	94.87	100.0	100.0
DL2 ACT3 01	100.0	100.0	100.0	100.0
DL2 ACT3 02	100.0	97.5	100.0	100.0
DL2 ACT3 03	80.0	97.29	100.0	100.0
DL2 ACT4 01	100.0	100.0	100.0	100.0
DL2 ACT4 02	100.0	100.0	100.0	100.0
DL2 ACT4 03	100.0	100.0	100.0	100.0
DL3 ACT01	100.0	100.0	100.0	100.0
DL3 ACT02	100.0	100.0	100.0	100.0
DL4 ACT01	100.0	96.15	100.0	100.0
DL4 ACT02	100.0	100.0	100.0	100.0
DL5 ACT01	31.57	61.85	91.66	83.33
DL5 ACT02	55.55	67.07	80.0	75.0

Table 5.

5.4.1. Comparison

Additionally, in order to evaluate our system and the contribution of the different components, our full system with, and without, the multiple context mechanism is compared to a baseline approach. The three systems under consideration are:

- **Baseline**: Basic system with no common-sense reasoning skills nor the multiple-context mechanism enabled. This is equivalent to a rule-based

Table 5: Description of the estimated sequence of events by our proposed system. Only those sequence with errors are displayed here, highlighted in bold, since the other ones are identical to Table 1. Erros are Code @f indicates the frame number at which the composite event starts.

Scene	**Actors**	**Frames**	**Composite Actions**
DL1 ACT3 01	1 M, **1 M**	1071	MBTS-C14 @f20, **MBTS-C6 @f298**, MEX @f665, FEX @f953
DL2 ACT3 02	**1 F**, 1 F	967	FBTS-C5 @f16, **FBTS-C6 @f224**, FCS-C3 @f380, FEX @f792, MEX @f847
DL5 ACT02	**3 M, 2 F**	1789	M1BTS-C17 @f18, **M2BTS-C12 @f294**,**F1BTS-C17 @f514**, **M2CS-C1 @f893**,M3B @f996, **M1CS-C2 @f1062**, **M2CS-C3 @f1168**, M1EX @f1354, M2EX @f1456, F1EX @f1542, **M3EX @f1712**
DL5 ACT01	3 M, 3 F	1432	F1BTS-C7 @f15, M1BTS-C14 @f179, F2BTS-C5 @f308, F1EX @f494, **M2BTS-C2** @f442, **M2CS-C17 @f610**, **F3BTS-C13 @f696**, M1EX @f818, **M3B @f799**, M2EX @f968, F2EX @f1029, F3EX @f1235, M3EX @f1274

systems that employs the same rules, world and domain knowledge as our proposed system. Action associations are carried out based on passenger boarding order and distances to the event location.

- **No multiple-context mechanism enabled**: This system is a different configuration of our proposed approach. It includes the HTN and common-sense mechanism, but its capabilities have been limited to only consider one context. Therefore, the multiple-context mechanism that supports the creation and maintenance of possible worlds is not available and only the most likely context at each time is preserved.

- **Full system**: This system exhibits the full functionality described in this paper.

The graphic in Figure 10 compares the performance of the three different systems under analysis for the four metrics considered here.

The full system outperforms the other two for all the four different metrics. However, the most significant improvement is that achieved for metric 4, which considers the whole passenger's story. This improvement demonstrates that corrections made when all atomic actions are put in perspective have an important impact on the obtained accuracy. The metric 4 accuraccy also demonstrates that the correct association of sensor events, or atomic events, may not be enough to fully understand the behaviour of the passengers in a sequence. Since the correctness of whole stories is what matters most for automatic surveillance, our proposal exhibits excellent potential for behaviour reasoning on public transport. It is also noticeable how the multiple-context mechanism allows results in a significant improvement, specially when creating the whole story. This is due to the fact that different hypotheses are preserved until all the information is available to create a coherent story and take the correct decision, instead of discarding hypotheses prematurely.

It is also worth mentioning the difference in performances obtained for metric 1. This metric indicates that the sensor event association is the most complex problem to solve, given the high uncertainty of the tracking video-analytics. An

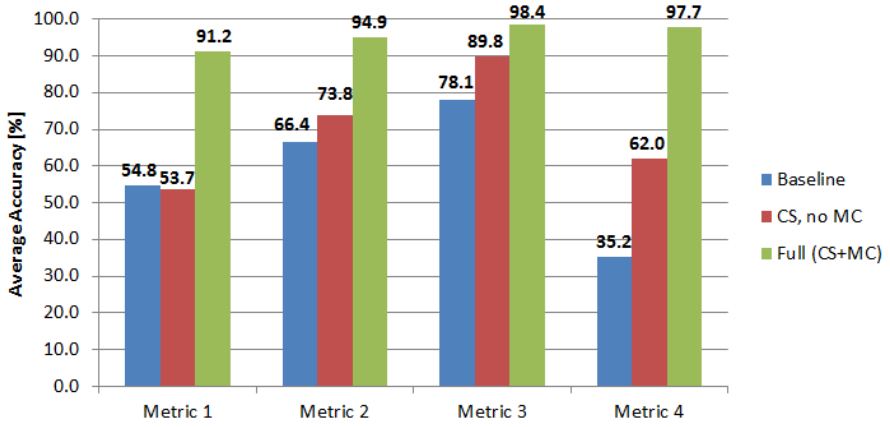

Figure 10: Average accuracy rates

analysis of the results demonstrates the importance of common-sense rules and the use of possible worlds. In this sense, common-sense rules have leveraged assertions, such as the fact that a person cannot be in two different places at the same time.

6. Conclusions

This paper describes an scalabe and distributed intelligent system for automatic surveillance and multiple-passenger behaviour monitoring in public-transport platforms based on heterogeneous and ambiguous sensor events and a common-sense reasoning approach. Sensor lack of precision, noise, uncertainty, and the presence of more than one passenger that makes event association difficult, are addressed by first homogenising the data and then providing causal explanations to the sensed events. Contrary to solutions that are provided with patterns or rules that describe human behaviour, our double-stage system is intended to give a causal explanation to the gathered events. A first stage associate the sensed atomic actions to passengers, whereas the second stage addresses the coherence and plausibility of the associations. The system makes use of common-sense reasoning, multiple-context consideration and hierarchical

association to provide the most likely final explanation to the behaviour of the passengers onboard.

As main theoretical contributions, this paper has introduced the use of the possible-world theory as a mechanism to handle the uncertainty and ambiguity of certain sensor events. Different *worlds* are created to track each of the possible scenarios in which such events are plausible, delaying the selection of just one until further information is available or until it cannot be delayed any longer. Moreover, a Hierarchical Task Network (HTN) planner is proposed as a mechanism to resemble the rationality that leads human behaviour. The planner has been theoretically formalised and empirically evaluated.

Our methodology is validated in a simulated bus scenario involving a variable number of passengers in different situations. Our system outperforms, over all four evaluation metrics, the rule-based baseline system which is provided with the same information, rules and knowledge. The greatest improvement was obtained when evaluating the correct whole-stories interpretations, which validates our capability to correct for the lack of common sense when whole scenarios are put in perspective. A main disadvantage to our approach, is the use of intrinsically ambiguous video sensors, which, while information rich, may produce a wrong explanation when the uncertainty increases due to growing scene complexity. However, even in the worst-case tested scenario, a 75% accuracy rate is obtained, outperforming the rule-based approach. It is also important to note that dangerous behaviour tends to occur at night while buses are mostly empty rather than overcrowded.

The main advantages of the proposed methodology are its scalability due to its distributed implementation, its ability to effectively combine a variety of heterogeneous rich and ambiguous sensors, the capacity to provide correct casual explanation under the presence of inconsistent and contradictory stories, and the avoidance of requiring accurate quantification of the uncertainty of sensors and events to provide valid explanations.

As a disadvantage, the use of intrinsically ambiguous video sensors, while rich in information, may produce wrong explanation when the uncertainty increases

due to a growing number of passengers and multitude of events take place in a spatio-temporal proximity. However, even in the worst-case tested scenario, a 75% accuracy rate is obtained, outperforming the rule-based approach. It is also important to note that most dangerous behaviours take place at night and in mostly empty busses rather than in overcrowded conditions.

Future work will address the drop in accuracy rate as the number of passengers increases. Since this is mainly caused by errors in the tracking system, the integration of better tracking algorithms with additional cameras should improve performance with larger passenger numbers. Secondly, human social behaviour in public-transport platforms will be modelled in collaboration with sociologists and incorporated into our knowledge base. Some patterns of social and anti-social behaviour have already been identified, but more interdisciplinary effort is required. Finally, we will extend the recordings, data capture and evaluation to real environments using actual busses and transport platforms.

Acknowledgement

This work has been partly funded by the Spanish Ministry of Economy and Competitiveness under project REBECCA (TEC2014-58036-C4-1-R) and by the Regional Government of Castilla-La Mancha under project SAND (PEII_2014_046_P). H. Zhou has been supported in part by UK EPSRC under grants EPH0496061, EPN5086641 and EPN0110741.

References

Artikis, A., Sergot, M., & Paliouras, G. (2010). A logic programming approach to activity recognition. In *Proceedings of the 2Nd ACM International Workshop on Events in Multimedia* EiMM '10 (pp. 3–8). New York, NY, USA: ACM.

Bae, I.-H. (2014). An ontology-based approach to {ADL} recognition in smart homes. *Future Generation Computer Systems*, *33*, 32 – 41. Special Section

on Applications of Intelligent Data and Knowledge Processing Technologies; Guest Editor: Dominik lzak.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004). *OWL Web Ontology Language Reference*. Technical Report W3C http://www.w3.org/TR/owl-ref/.

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *Image and Video Processing*, *2008*, 1–10.

Bourdev, L., & Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *In Proc. 12th International Conference in Computer Vision*.

Candan, K. S., Liu, H., & Suvarna, R. (2001). Resource description framework: Metadata and its applications. *SIGKDD Explor. Newsl.*, *3*, 6–19.

Chaaraoui, A. A., Climent-Perez, P., & Florez-Revuelta, F. (2012). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, *39*, 10873 – 10888.

Chen, W., & Fahlman, S. E. (2008). Modelling mental context and their interactions. In *In AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*.

Cilla, R., Patricio, M. A., Berlanga, A., & Molina, J. M. (2012). A probabilistic, discriminative and distributed system for the recognition of human actions from multiple views. *Neurocomputing*, *75*, 78 – 87. Brazilian Symposium on Neural Networks (SBRN 2010) International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010).

Cocea, M., & Magoulas, G. D. (2012). User behaviour-driven group formation through case-based reasoning and clustering. *Expert Systems with Applications*, *39*, 8756 – 8768.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, *60*, 685–700.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *NUMERISCHE MATHEMATIK*, *1*, 269–271.

Divers, J. (2002). *Possible worlds. Problems of philosophy*. Routledge.

Do, T., Loke, S., & Liu, F. (2013). Healthylife: An activity recognition system with smartphone using logic-based stream reasoning. In K. Zheng, M. Li, & H. Jiang (Eds.), *Mobile and Ubiquitous Systems: Computing, Networking, and Services* (pp. 188–199). Springer Berlin Heidelberg volume 120 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*.

Edwards, C. (2014). Decoding the language of human movement. *Commun. ACM*, *57*, 12–14.

Fahlman, S. (2011). Using scone's multiple-context mechanism to emulate human-like reasoning.

Fahlman, S. E. (2006). Marker-Passing Inference in the Scone Knowledge-Base System. In *First International Conference on Knowledge Science, Engineering and Management (KSEM'06)*. Springer-Verlag (Lecture Notes in AI).

Fahlman, S. E. (2010). The Scone knowledge-base project. Available online at: `http://www.cs.cmu.edu/~sef/scone/`. Retrieved on February 28th, 2015.

Gomez, M. J., Garcia, F., Martin, D., de la Escalera, A., & Armingol, J. M. (2015). Intelligent surveillance of indoor environments based on computer vision and 3d point cloud fusion. *Expert Systems with Applications*, *42*, 8156 – 8171.

Gomez-Romero, J., Patricio, M. A., Garcia, J., & Molina, J. M. (2011). Ontology-based context representation and reasoning for object tracking and

scene interpretation in video. *Expert Systems with Applications*, *38*, 7494 – 7510.

Han, S.-G., Lee, S.-G., & Jo, G.-S. (2005). Case-based tutoring systems for procedural problem solving on the www. *Expert Systems with Applications*, *29*, 573 – 582.

Hogg, C., Kuter, U., & Muñoz-Avila, H. (2009). Learning Hierarchical Task Networks for Nondeterministic Planning Domains. In *Twenty-First Internat. Joint Conf. on Artificial Intelligence (IJCAI-09)*.

Hong, X., Huang, Y., Ma, W., Varadarajan, S., Miller, P., Liu, W., Jose Santofimia Romero, M., Martinez del Rincon, J., & Zhou, H. (2016). Evidential event inference in transport video surveillance. *Computer Vision and Image Understanding*, *144*, 276–297.

Kuipers, B. (1994). *Qualitative Reasoning, Modelling and Simulation with Incomplete Knowledge*. MIT Press.

Li, Y., Huang, C., & Nevatia, R. (2009). Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Ltd, V. M. S. (1984). Vicon motion capture. URL: `http://www.vicon.com/`.

McCarthy, J. (1968). Programs with common sense. In *Semantic Information Processing* (pp. 403–418). volume 1.

McLaughlin, N., del Rincon, J. M., & Miller, P. (2014). Dense multiperson tracking with robust hierarchical linear assignment. *IEEE Transactions on Cybernetics*, .

Minsky, M. (1999). The emotion machine: from pain to suffering. In *Creativity & Cognition* (pp. 7–13).

Mueller, E. T. (2006). *Commonsense Reasoning*. Morgan Kaufmann.

Mueller, E. T. (2007). Understanding goal-based stories through model finding and planning. In *Intelligent Narrative Technologies, Papers from the 2007 AAAI Fall Symposium* (pp. 95–102).

Nebel, J., Lewandowski, M., Thevenon, J., Martinez, F., & Velastin, S. (2011). Are current monocular computer vision systems for human action recognition suitable for visual surveillance applications? In *In International Symposium on Visual Computing*.

del Rincon, J. M., Santofimia, M. J., & Nebel, J.-C. (2013). Common-sense reasoning for human action recognition. *Pattern Recognition Letters*, *34*, 1849 – 1860. Smart Approaches for Human Action Recognition.

Rodriguez, N. D., Cuellar, M. P., Lilius, J., & Calvo-Flores, M. D. (2014). A fuzzy ontology for semantic modelling and recognition of human behaviour. *Knowledge-Based Systems*, *66*, 46 – 60.

SanMiguel, J. C., & Martinez, J. M. (2012). A semantic-based probabilistic approach for real-time video event recognition. *Computer Vision and Image Understanding*, *116*, 937 – 952.

Santofimia, M. J., del Rincon, J. M., & Nebel, J.-C. (2014). Episodic reasoning for vision-based human action recognition. *Scientific World Journal*, *2014*.

Sebbak, F., Chibani, A., Amirat, Y., Mokhtari, A., & Benhammadi, F. (2013). An evidential fusion approach for activity recognition in ambient intelligence environments. *Robotics and Autonomous Systems*, *61*, 1235 – 1245. Ubiquitous Robotics.

Stewart, D., Wang, H., Shen, J., & Miller, P. (2009). Investigations into the robustness of audio-visual gender classification to background noise and illumination effects. In *Digital Image Computing: Techniques and Applications, 2009. DICTA '09.* (pp. 168–174).

Viola, P., & Jones, M. (2014). Robust real-time face detection. *International Journal of Computer Vision*, *57*, 137 – 154.

Wilensky, R. (1983). *Planning and Understanding*. Reading, MA: Addison-Wesley.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford scholarship online. Oxford University Press.